基于改进Apriori关联规则算法的信令分析

Signaling Analysis Based on Improved Apriori Association Rule Algorithm

唐学军，周达谋，李慧莲（中国联通广东分公司，广东 广州 510627）
Tang Xuejun，Zhou Damou，Li Huilian（China Unicom Guangdong Branch，Guangzhou 510627，China）

摘　要：
传统信令分析方法需要专业人员找出可能与失败码相关的聚集的信令字段值或值组合及其导致失败的概率，并定位网络问题，操作复杂且效率低。通过改进Apriori关联规则算法，将探寻聚集的字段值或其组合的过程转换成发现失败码和相关信令字段值的关联规则。在计算频繁项集时，通过设置最小支持度阈值找出包含失败码的频繁项，将待分析失败码作为后项，减少了算法的复杂度和算力要求，并通过置信度和提升度找出与后项强关联的属性，实现了对失败码集中属性的快速高效识别。

关键词：
信令分析；关联规则；Apriori算法
doi：10.12045/j.issn.1007-3043.2024.09.011
文章编号：1007-3043（2024）09-0063-05
中图分类号：TN915
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Traditional signaling analysis methods require professionals to identify clustered signaling field values or combinations of values that may be related to failure codes and the probability of causing failure codes. Based on this，network problems can be located，which is complex and inefficient to operate. It improves the Apriori association rule algorithm to transform the process of exploring aggregated field values or their combinations into association rules for discovering failure codes and related signaling field values. When calculating the frequent itemset，the minimum support threshold is set to identify the frequent items containing failure codes. The failure codes to be analyzed are treated as the subsequent items，which reduces the complexity and computational power requirements of the algorithm，and the attributes strongly associated with the subsequent items are identified through confidence and enhancement，achieving fast and efficient identification of attributes in the failure code set.

Keywords：
Signaling analysis；Association rules；Apriori algorithm

引用格式：唐学军，周达谋，李慧莲. 基于改进Apriori关联规则算法的信令分析[J]. 邮电设计技术，2024(9)：63-67.

1 概述

信令记录是通信网络中用户与网络交互的主要信息源，包含了许多关键的操作参数和状态信息。传统的信令分析方法通常通过对信令记录的某些字段进行深入分析，找出产生失败码的集中属性，进而定位和诊断网络问题。信令记录一般会包括用户相关信息（如MSISDN、IMSI）、终端相关信息（如终端IMEI、终端TAC等）、设备相关信息（如设备编码等）和结果相关信息（如成功、失败及失败码等）等字段，信令分析的主要目的是找出失败原因，定位出与失败相关的字段信息（如某些号码段、某些终端或某个设备）。

针对某个失败码的分析，传统的做法是：第1步，先筛选包含该失败码的记录，从中选取对失败码具有潜在关联的特定字段或字段组合（如用户号码、终端类型、接入设备、APN、核心网设备编码等），然后计算所选特定字段的值或字段组合的值组合是否有聚集趋势，最后获取有聚集趋势的字段值或值组合；第2

收稿日期：2024-08-09

步，计算所选字段值或者值组合是否与该失败码强关联，即在全部记录中查询包含所选字段值或值组合的记录，计算该失败码发生的概率，概率越高，说明所选字段值或值组合越可能是导致该失败码的主要原因。这个过程中最困难的是字段组合的选择，选择正确时可以很快找到与失败相关的值组合，从而发现问题，但是由于信令记录字段众多，字段组合的复杂度高，而且不同字段的值变化多样，导致目标失败码的值组合很难被发现，需要丰富的专家经验进行判断。对比数据挖掘中的关联规则算法[1-2]，可以发现这个过程本质上就是发现信令记录中含失败码的强关联规则。本文从Apriori关联规则算法出发，简单阐述了算法逻辑，并结合实际信令分析过程进行算法改进，从而实现高效自动化的信令分析。

2 关联规则算法

2.1 关联规则相关指标

关联规则算法是在一个由一系列元素组合构成的集合D中，找到元素X和Y同时出现的情况$X\rightarrow Y$，$X\rightarrow Y$即为关联规则。在D中，X、Y同时出现的概率称为支持度，在D包含X的子集中Y也出现的概率称为置信度（见表1）。置信度与Y在D中出现的概率之比称为提升度，提升度反映了X、Y的相关性，提升度等于1表示X、Y没有相关性，提升度小于1表示X和Y负相关。如果某个关联规则$X\rightarrow Y$满足设定的最小支持度阈值和最小置信度阈值且提升度满足某个大于1的阈值，则认为关联规则是有意义和强相关的。

表1 关联规则算法相关指标定义

指标	定义	意义
支持度	X、Y同时出现数/总数	支持度越高，X、Y同时出现的概率越大
置信度	X、Y同时出现数/X出现数	置信度越高，X出现的情况下出现Y的概率越高
提升度	置信度与Y在全集中出现的概率之比	提升度反映了X、Y的相关性，提升度大于1时，数值越大，X、Y相关性越强

2.2 Apriori关联规则算法简述

Apriori关联规则算法是一种经典的挖掘关联规则频繁项集的算法，它利用逐层搜索的迭代方法找出数据库中项集的关系以形成规则，该过程由连接与剪枝组成。该算法中项集即为项的集合，包含K个项的集合为k项集，如牛奶、面包组成一个集合{牛奶、面包}，其中牛奶和面包为项，{牛奶、面包}为项集，该项集被称为2项集。所有支持度大于最小支持度阈值的项集称为频繁项集，频繁项集的子集必为频繁项集，非频繁项集的超集一定是非频繁的，在具体计算过程中只需要关注频繁项集。Apriori关联规则算法实现过程为：首先，找出频繁1项集，再从频繁1项集中增加元素产生频繁2项集，并不断迭代找出所有的频繁项集（见图1），然后从频繁项集中找出置信度不小于最小置信度阈值且提升度不小于最小提升度阈值的关联规则，最终得到满足最小支持度、最小置信度和最小提升度要求的规则，即具有强关联性的$X\rightarrow Y$（见图2）。

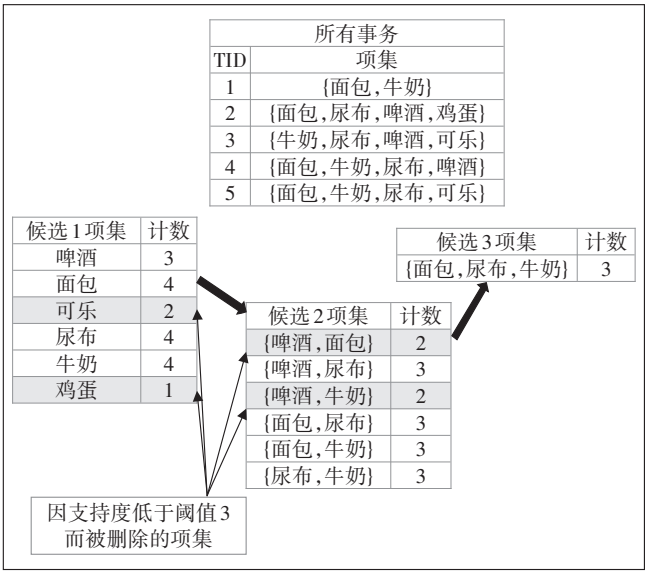

图1 使用Apriori关联规则算法产生频繁项集的过程

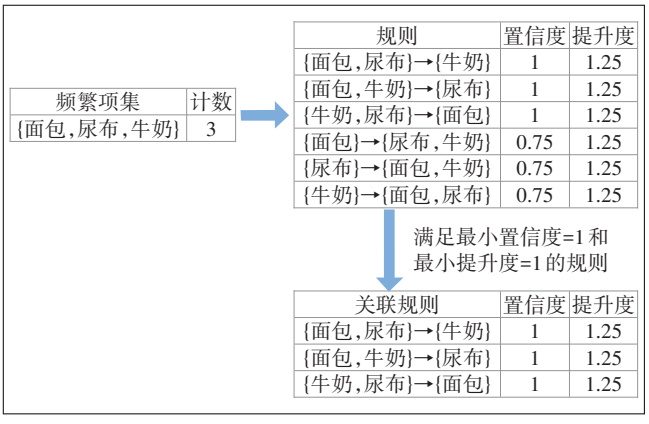

图2 使用Apriori关联规则算法产生关联规则的过程

3 基于改进Apriori关联规则算法的信令分析

3.1 传统信令分析流程及与Apriori关联规则算法的关系

Apriori关联规则算法挖掘过程主要包含2个阶段：第1阶段先从集合中找出所有满足最小支持度阈

值的频繁项集，第2阶段再从这些频繁项集中产生关联规则，计算其置信度，然后保留那些置信度大于等于最小置信度阈值且提升度大于最小提升度阈值的关联规则。对于传统信令分析方法，第1步从包含失败码的记录中发掘具有聚集趋势的字段值或者值组合，相当于关联算法中根据设置的最小支持度发现频繁项集；第2步，结合所有信令记录筛选出上一步中与失败码强相关的字段值或者值组合，相当于在频繁项集中根据置信度和提升度确定规则（见图3）。

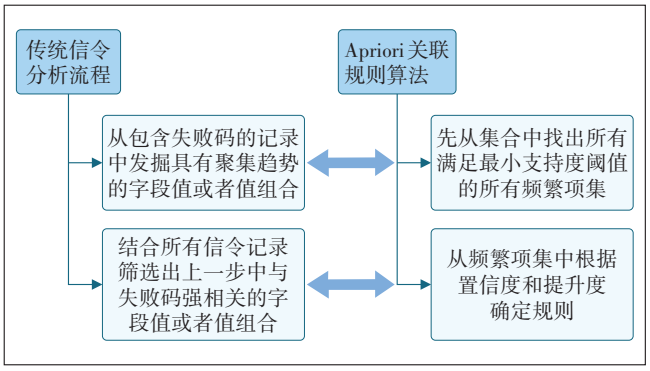

图3　传统信令分析流程及与Apriori关联规则算法的关系

3.2 基于信令数据的Apriori算法改进

将Apriori关联规则算法用于信令分析时，首先需要确定选择哪些信令记录进行分析。每一条信令记录就是一个事务，本文将信令过程结果为成功和待分析失败码的记录组成的信令记录集作为数据集，由于目标是找出与失败码相关的字段信息，所以只需要关注这些字段信息导致信令过程成功或出现该失败码的可能性，因此只需要选择这2种记录作为数据集，就可以找出产生失败码的集中字段值并定位网络问题。考虑到绝大多数情况下信令记录中结果为成功的记录数远大于失败的记录数，包含成功记录的数据集数量会很大，需采取相关举措提升算法效率。

首先，本文将信令中各字段值作为Apriori关联规则算法的项。虽然组合复杂，但在信令记录中，同一字段的不同字段值组合一定不是频繁项集，只有不同字段的值组合才会成为频繁项集，这样可大大减少频繁项集的计算量。其次，本文关注的是产生失败码的规则，因此失败码作为后项是频繁项集的必要元素。本文使用的关联规则算法不是从1项集而是从2项集开始产生频繁项集，当2项集中包含待分析失败码的字段值组合时，该项集才会成为频繁项集，相当于只对信令过程结果为待分析失败码的记录计算频繁项集，结合最小支持度阈值，大大减少了频繁项集的计算量。此外，在信令分析中，选择结果为成功的信令记录主要用于统计支持度和提升度，不参与频繁项集的获取，这样大大提升了算法的效率。

分析信令失败码的目的是找出该失败码出现的主要原因。项集支持度等于项集事务数除以总事务数，根据经验，建议最小支持度阈值设置为待分析失败码记录数占总记录数比例的50%以上。置信度的设置与所选字段值对失败码的定位准确性有关，项集置信度等于项集事务数除以待分析失败码记录数，根据经验，建议最小置信度阈值设置为80%以上。规则的提升度等于项集的置信度除以待分析失败码记录数占总事务数比例，提升度一般设置为大于1。

改进的Apriori关联规则算法的实现过程是先从信令记录中选择信令过程结果为成功和待分析失败码的记录作为事务，每个记录中用户、终端和网络相关信息等各个字段的各种字段值作为前项X，待分析的结果相关信息（如UE ESM原因=27）作为后项Y，每个前项与后项组成2项集，计算满足最小支持度的频繁项集（高频项目组），再将频繁项集中的前项两两组合，每个组合与后项组成3项集，计算满足最小支持度的频繁项集，不断重复，直到不再产生新的频繁项集。对所有频繁项集，计算规则前项到后项（$X \rightarrow Y$）的置信度和提升度，满足最小置信度和提升度的最大集合即为目标关联规则，每条规则对应的前项就是要分析的结果产生的关联属性（见图4）。通过对结果相关信息的遍历，可以得到所有失败码的关联字段值。

4 实证分析

对某个时间段20万条S1MME接口的ATTACH过程的信令记录进行分析，选取成功记录及待分析失败码记录的集合，最小支持度设置为该后项记录占总记录数的80%，最小置信度设置为80%，提升度设置为2。信令记录包括了序号、开始时间、结束时间、流程类型、流程状态、IMSI、MSISDN、IMEI、IMEI软件版本、MS IP、接入网类型、路由/跟踪区编码、小区编码、小区名称、工作模式、APN、eNodeB名称、eNodeB信令面IP、MME名称、MME IP、终端品牌、终端型号、EMM事务、EMM原因、UE ESM事务、UE ESM原因、网络ESM事务、网络ESM原因、S1AP事务、S1AP原因、身份识别原因、鉴权原因、安全模式原因、CSFB业务指示、协商成功的上行最大速率（kbit/s）、协商成功的下行最大速率

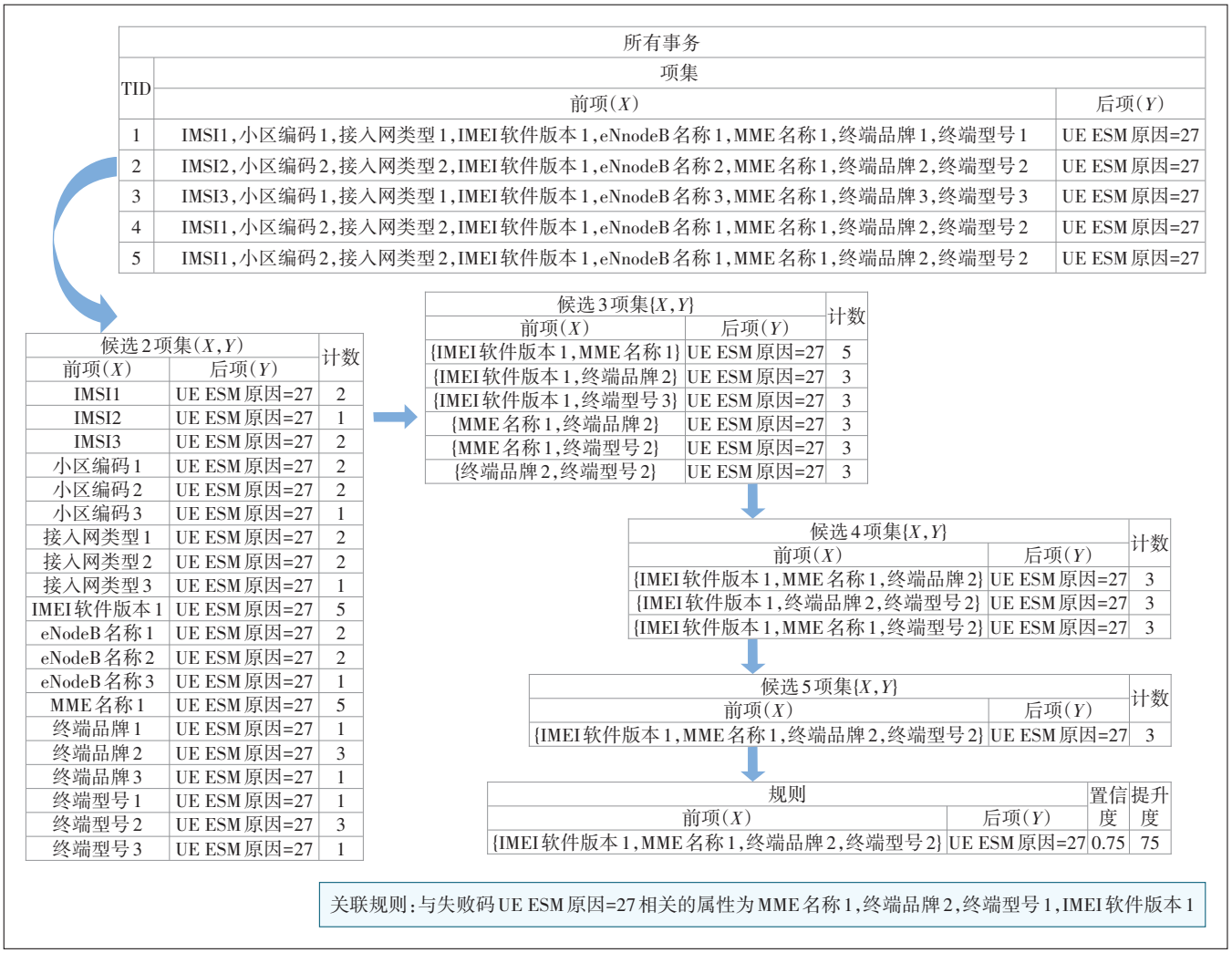

图4 基于信令数据的Apriori算法的改进过程

(kbit/s)、协商成功的QoS级别标识符、协商成功的分配保持优先级、用户IP地址类型等39个字段。

4.1 数据预处理

在对数据进行清洗后，为了信令分析的准确性，还要从现有字段中生成新的包含重要信息的字段，包括但不限于以下内容。

a) 数据离散化。把时间字段(开始时间、结束时间)按一定的时间间隔(如小时或分钟)进行分段，生成时间分段字段，以判断失败是否与时段相关。

b) 将号码字段(IMSI、MSISDN)分段。IMSI选择前10个字符串生成IMSI前缀字段，MSISDN取前7个字符串生成MSISDN前缀字段，以判断失败是否与用户所属区域相关。

c) 将IMEI分段。IMEI取前8个字符串，即终端的TAC号，标识了终端所属厂家及型号，以判断失败是否与终端厂家及型号相关。

4.2 选择信令分析目标

以字段UE ESM原因、网络ESM原因的每个选项作为分析目标，以流程状态='成功'的记录、合并分析目标相关记录作为数据集，选择IMSI、MSISDN、IMEI、IMEI软件版本、MS IP、接入网类型、路由/跟踪区编码、小区编码、小区名称、工作模式、APN、eNodeB名称、eNodeB信令面IP、MME名称、MME IP、终端品牌、终端型号、时间分段字段、IMSI前缀字段、MSISDN前缀字段、TAC字段作为关联字段，分析关联字段与目标字段的关联关系。

4.3 生成关联规则

对于目标字段的每个选项，先计算每个关联字段的每个值与目标字段值的组合记录数是否符合支持度要求，对于符合支持度要求的值，计算其两两组合+

目标字段该选项的记录是否符合支持度要求。由于每个字段只有1个选项值，因此不会出现同一个字段的2个选项的组合满足支持度要求的情况。对于满足支持度要求的2个前项组合，计算组合为3个组合+目标字段该选项的记录是否符合支持度要求，直到找出所有符合支持度要求的组合或达到预设的最大前项数。本例找出了所有的符合支持度要求的组合，并对目标字段的每个选项的符合支持度要求的组合按置信度和提升度进行排序，具体分析结果如表2所示。

4.4 分析结果说明

表2 基于改进Apriori算法关联规则的信令分析结果

失败码	失败记录数	关联规则	关联规则相关失败记录数	支持度/%	置信度/%	提升度
UE ESM原因 = 27 Missing or unknown APN	19 072	终端品牌：QUECTEL；终端型号：EC20-CE；IMEI软件版本：10，APN：UNIM2M.xxM2MAPN	17 852	9.11	99.72	10
UE ESM原因 = 997 UE Context Release	1 394	无	–	–	–	–
UE ESM原因 = 50 PDN type IPv4 only allowed	1 200	MSISDN：155xxxx0705；终端型号：SHARK KLE-A0；IMSI：46001xxxx677435；路由/跟踪区编：46001xx34；小区编码：46001416B282；小区名称：46001xx6B282；eNodeB信令面IP：xx.72.89.141；IMEI：866475042xxxx63	1 197	0.67	100	148
UE ESM原因 = 998 Procedure Terminated	551	无	–	–	–	–
UE ESM原因 = 29 User authentication failed	421	APN：xxxx.JLREXT.COM	419	0.24	81.51	343
UE ESM原因 = 51 PDN type IPv6 only allowed	113	MSISDN：132xxxx5375；IMEI：868358020xxxx46；eNodeB信令面IP：x.x.16.21；终端型号：MILAI M8；小区名称：46001xx9B403；终端品牌：MILAI；小区编码：46001xx9B403；IMSI：46001xxxx361333	106	0.06	100	1 567
网络ESM原因 = 997 UE Context Release	1 030	无	–	–	–	–
网络ESM原因 = 31 Request rejected，unspecified	109	无	–	–	–	–
网络ESM原因 = 1000 Timeout	349	无	–	–	–	–

失败码UE ESM原因 = 27 Missing or unknown APN，经确认是由某终端型号某个版本的终端在进行物联网业务时引起的，属于终端版本问题，是该终端版本的APN配置错误引起的。

失败码UE ESM原因 = 50 PDN type IPv4 only allowed，经确认是某用户终端在发起PDN建立请求时PDN类型为IPv6，而网络侧用户配置只支持IPv4导致的。

失败码UE ESM原因 = 29 User authentication failed，经确认是接入某个专有APN的用户群开户模板有问题，导致用户鉴权失败，需要核查用户开户模板。

失败码UE ESM原因 = 51 PDN type IPv6 only allowed，经确认是某用户终端在发起PDN建立请求时PDN类型为IPv4，而网络侧用户配置只支持IPv6导致的。

其他失败码没有发现有强聚集特性。

5 总结

改进的关联算法有针对性地将信令记录中待分析的结果作为后项，将选择的字段值作为前项，并将两者组成集合进行频繁项集计算，在满足设定的最小置信度和提升度阈值的前提下找出符合要求的规则。在监控到信令指标异常时，可通过改进的Apriori关联规则算法对该指标相关信令过程的每一个失败码进行自动分析，找出高度关联的字段值组合，从而快速发现和定位网络问题，极大地提升了信令分析和优化的效率。

参考文献：

[1] 熊平. 数据挖掘算法与Clementine实践[M]. 北京：清华大学出版社，2011：80-120.
[2] TAN P N，STEINBACH M，KUMAR V. 数据挖掘导论：完整版[M]. 北京：人民邮电出版社，2011：207.

作者简介：
唐学军，毕业于清华大学，高级工程师，主要从事移动网优化及客户感知分析工作；周达谋，毕业于中国科学院科技战略咨询研究院，主要从事移动网优化及客户感知分析工作；李慧莲，毕业于重庆邮电大学，正高级工程师，主要从事移动网优化及客户感知分析工作。